GroupH_HM2

Simonutti, Younes Pour Langaroudi, Billo, Tavano, Vicig

2024-12-04

Contents

FSDS - Chapter 4

Ex 4.2

For a sequence of observations of a binary random variable, you observe the geometric random variable (Section 2.2.2) outcome of the first success on observation number $y = 3$. Find and plot the likelihood function.

Ex 4.4

For the Students data file and corresponding population, find the ML estimate of the population proportion believing in life after death. Construct a Wald 95% confidence interval, using its formula (4.8). Interpret.

(4.8)

$$\hat{\pi} = z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Ex 4.38

For independent observations $y_1, ..., y_n$ having the geometric distribution (2.1):

(a) Find a sufficient statistic for π.
(b) Derive the ML estimator of π.

Ex 4.44

Refer to the previous two exercises. Consider the selling prices (in thousands of dollars) in the Houses data file mentioned in Exercise 4.31.

(a) Fit the normal distribution to the data by finding the ML estimates of μ and σ for that distribution.

(b) Fit the log-normal distribution to the data by finding the ML estimates of its parameters.

(c) Find and compare the ML estimates of the mean and standard deviation of selling price for the two distributions.

(d) Superimpose the fitted normal and log-normal distributions on a histogram of the data. Which distribution seems to be more appropriate for summarizing the selling prices?

Ex 4.54

Consider n independent observations from an exponential pdf $f(y;\lambda) = \lambda e^{-\lambda y}$ for $y \geq 0$, with parameter $\lambda > 0$ for which $E(Y) = \frac{1}{\lambda}$.

(a) Find the sufficient statistic for estimating λ.

(b) Find the maximum likelihood estimator of λ and of $E(Y)$.

(c) One can show that $2\lambda(\sum_i Y_i)$ has a chi-squared distribution with $df = 2n$. Explain why $2\lambda(\sum_i Y_i)$ is a pivotal quantity, and use it to derive a 95% confidence interval for λ.

FSDS - Chapter 5

Ex 5.68

Explain why the confidence interval based on the Wald test of $H_0 : \theta = \theta_0$ is symmetric around $\hat{\theta}$ (i.e., having center exactly equal to $\hat{\theta}$. This is not true for the confidence intervals based on the likelihood-ratio and score tests.) Explain why such symmetry can be problematic when θ and $\hat{\theta}$ are near a boundary, using the example of a population proportion that is very close to 0 or 1 and a sample proportion that may well equal 0 or 1.

FSDS - Chapter 6

Ex 6.12

For the UN data file at the book's website (see Exercise 1.24), construct a multiple regression model predicting Internet using all the other variables. Use the concept of multicollinearity to explain why adjusted R^2 is not dramatically greater than when GDP is the sole predictor. Compare the estimated GDP effect in the bivariate model and the multiple regression model and explain why it is so much weaker in the multiple regression model.

Ex 6.14

The data set Crabs2 at the book's website comes from a study of factors that affect sperm traits of male horseshoe crabs. A response variable, *SpermTotal*, is the log of the total number of sperm in an ejaculate. It has $\bar{y} = 19.3$ and $s = 2.0$. The two explanatory variables used in the R output are the horseshoe crab's *carapacewidth* (CW, mean 18.6 *cm*, standard deviation 3.0 *cm*), which is a measure of its size, and *color* (1 = dark, 2 = medium, 3 = light), which is a measure of adult age, darker ones being older.

Ex 6.30

When the values of y are multiplied by a constant c, from their formulas, show that s_y and $\hat{\beta}_1$ in the bivariate linear model are also then multiplied by c. Thus, show that r = $\hat{\beta}_1(\frac{s_x}{s_y})$ does not depend on the units of measurement.

Ex 6.42

You can fit the quadratic equation E(Y) = $\beta_0 + \beta_1 x + \beta_2 x^2$ by fitting a multiple regression model with $x_1 = x$ and $x_2 = x^2$.

(a) Simulate 100 independent observations from the model $Y = 40.0 - 5.0x + 0.5x^2 + \epsilon$, where X has a uniform distribution over [0, 10] and $\epsilon \sim N(0, 1)$. Plot the data and fit the quadratic model. Report how the fitted equation compares with the true relationship

(b) Find the correlation between x and y and explain why it is so weak even though the plot shows a strong relationship with a large R^2 value for the quadratic model.

Ex 6.52

F statistics have alternate expressions in terms of R^2 values.

(a) Show that for testing $H_0 : \beta_1 = ... = \beta_p = 0$,

$$F = \frac{(TSS - SSE)/p}{SSE/[n - (p + 1)]}$$

is equivalent to:

$$F = \frac{R^2/p}{(1 - R^2)/[n - (p + 1)]}$$

Explain why larger values of R^2 yield larger values of F.

(b) Show that for comparing nested linear models,

$$F = \frac{(SSE_0 - SSE_1)/(p_1 - p_0)}{SSE_1/[n - (p_1 + 1)]} \quad = \quad \frac{R_1^2 - R_0^2/(p_1 - p_0)}{(1 - R_1^2)/[n - (p_1 + 1)]}$$

FSDS - Chapter 7

Ex 7.4

Analogously to the previous exercise, randomly sample 30 X observations from a uniform in the interval (-4,4) and conditional on X = x, 30 normal observations with E(Y) = $3.5x^3 - 20x^2 + 0.5x + 20$ and σ = 30. Fit polynomial normal GLMs of lower and higher order than that of the true relationship. Which model would you suggest? Repeat the same task for E(Y) = $0.5x^3 - 20x^2 + 0.5x + 20$ (same σ) several times. What do you observe? Which model would you suggest now?

Ex 7.20

In the Crabs data file introduced in Section 7.4.2, the variable y indicates whether a female horseshoe crab has at least one satellite (1 = yes, 0 = no).

(a) Fit a main-effects logistic model using weight and categorical color as explanatory variables. Conduct a significance test for the color effect, and construct a 95% confidence interval for the weight effect.

(b) Fit the model that permits interaction between color as a factor and weight in their effects, showing the estimated effect of weight for each color. Test whether this model provides a significantly better fit.

(c) Use AIC to determine which models seem most sensible among the models with (i) interaction, (ii) main effects, (iii) weight as the sole predictor, (iv) color as the sole predictor, and (v) the null model.

Ex 7.26

A headline in The *Gainesville Sun* (Feb. 17, 2014) proclaimed a worrisome spike in shark attacks in the previous two years. The reported total number of shark attacks in Florida per year from 2001 to 2013 were 33, 29, 29, 12, 17, 21, 31, 28, 19, 14, 11, 26, 23. Are these counts consistent with a null Poisson model? Explain, and compare aspects of the Poisson model and negative binomial model fits.